Curso Machine Learning

Tema 4: Reducción y Agrupamiento

Magdalena Lucini, Luis Duarte, Sebastián Filipigh

FaCENA - UNNE - 2023

Contenidos

Ejemplo 1

- Datos
  - $n = 26$ individuos (países)
  - $p = 10$ variables (indicadores demográficos)
- cf. Population Reference Bureau (http://www.prob.org/)

País	Tasa de Nacimientos(%)	Tasa Mortalidad(%)	...	Población
Afganistán	47	21		6384000
Albania	13	21		1443000
Argentina	19	8		36324000
⋮				
Zimbabwe	31	21		5024000

Table: Ejemplo 1 de datos multivariados

Ejemplo 2

- Datos
  - $n = 507$ individuos (personas)
  - $p = 24$ variables (indicadores del cuerpo)

Persona	Prof. Pecho (cm)	Largo Pierna(cm)	...	Edad	Peso (kg)
1	17.7	106.2		21	65.6
2	16.9	110.5		23	71.8
⋮					
506	15.5	107.1		33	66.4
507	20.4	100.5		38	57.3

Table: Ejemplo 2 de datos multivariados

Preguntas

- Extraer y sintetizar variables pertinentes:
  - ej 1→ ¿hay indicadores similares?
  - ej 2→ ¿hay indicadores de cuerpos parecidos?
- Formar grupos de individuos con mismas características:
  - ej 1→ ¿hay países que se comporten de manera similar?
  - ej 2→ ¿hay personas con características corporales similares?
- Modelar una variable en función de otras:
  - ej 1→ ¿puede explicarse la tasa de mortalidad en función de las otras variables medidas?
  - ej 2→ ¿podemos explicar el peso de una persona?

Notación

individuos	Variable 1	...	Variable j	...	Variable p
1			$\vdots$		
$\vdots$			$\vdots$		
i	...	...	$x_{i,j}$		
$\vdots$					
n					

Table: Representación esquemática de una tabla de datos multivariados

- n : número de individuos
- p: número de variables
- $x_{i,j}$: respuesta de un individuo/objeto/elemento i a la variable j

Reducción de dimensionalidad

Objetivo

Reducir la dimensionalidad de los datos: describir un conjunto de datos con un número menor de variables

Métodos:

- Reducir características (variables): por "intuición", eliminar las que tienen poca varianza, eliminar características redundantes.
- **Análisis de Componentes Principales (ACP) variables cuantitativas**
- Análisis de Correspondencias variables cualitativas
- Embeddings

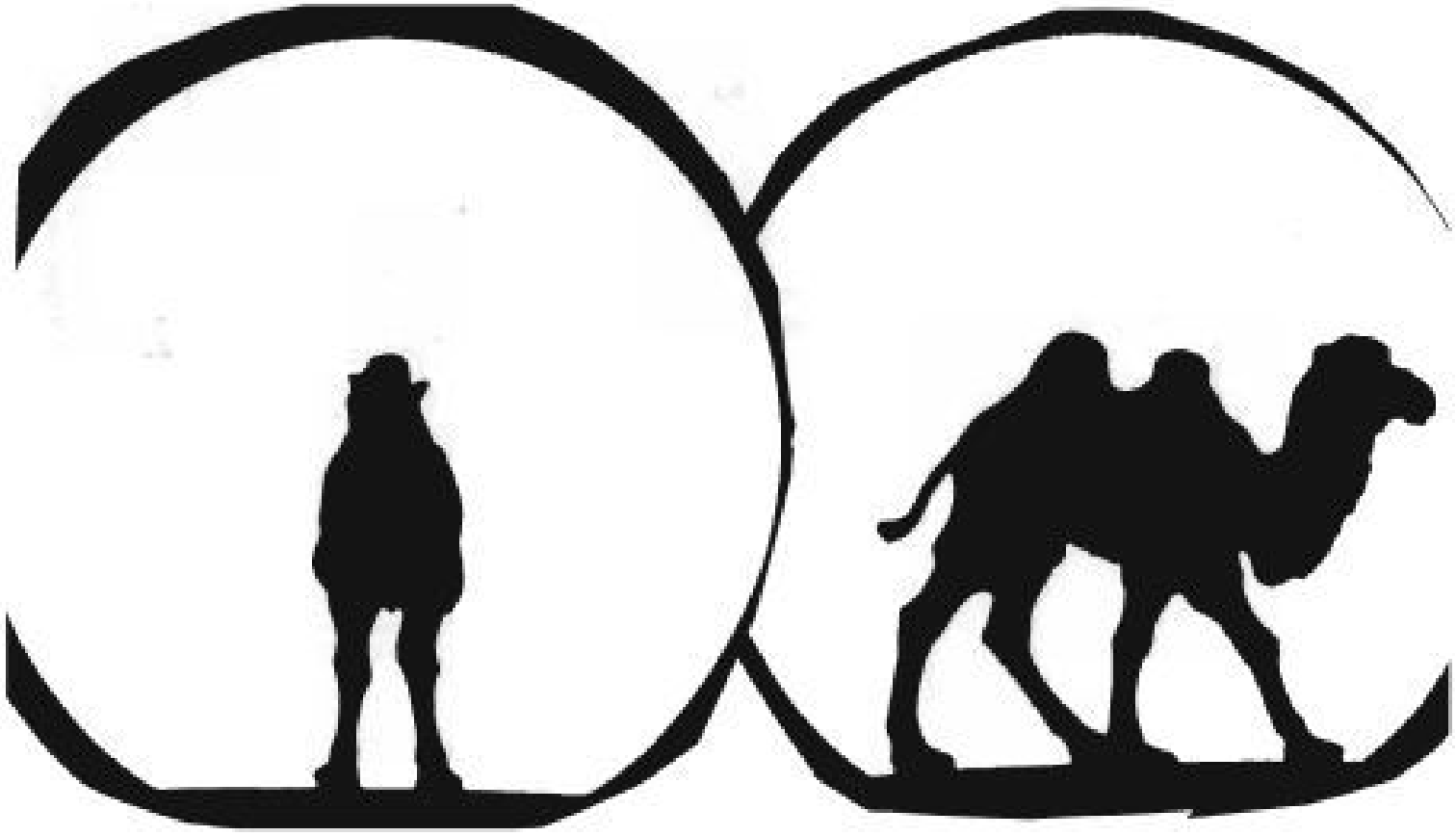
Aplicaciones

compresión de datos, reconstrucción de datos, preprocesamiento de datos (antes de agrupamiento), etc..

Análisis de Componentes Principales - ACP

Objetivo

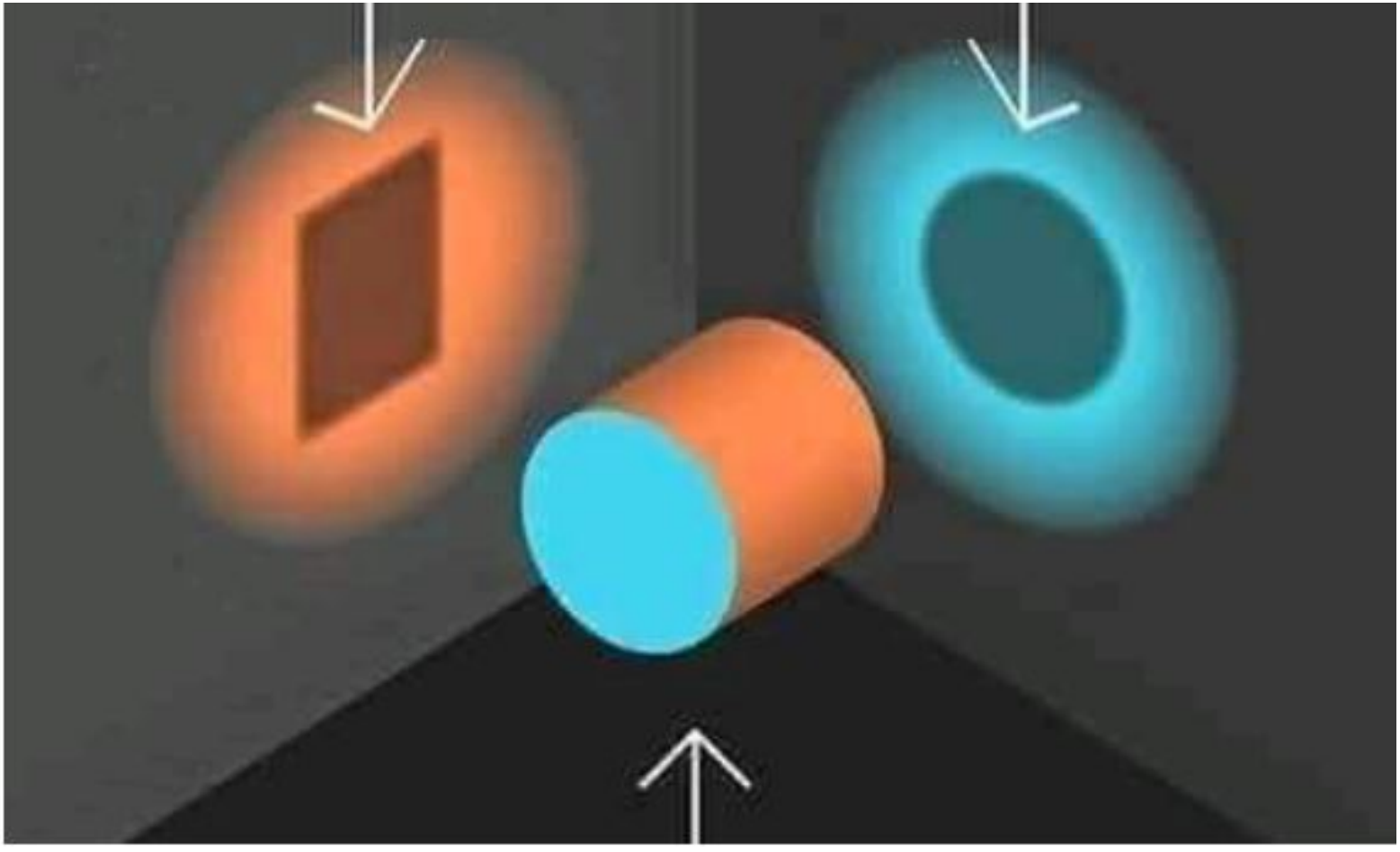
Encontrar el subespacio que mejor describa los datos: El más "cercano" por proyecciones.



Análisis de Componentes Principales - ACP

Objetivo

Encontrar el subespacio que mejor describa los datos, teniendo en cuenta el propósito del estudio



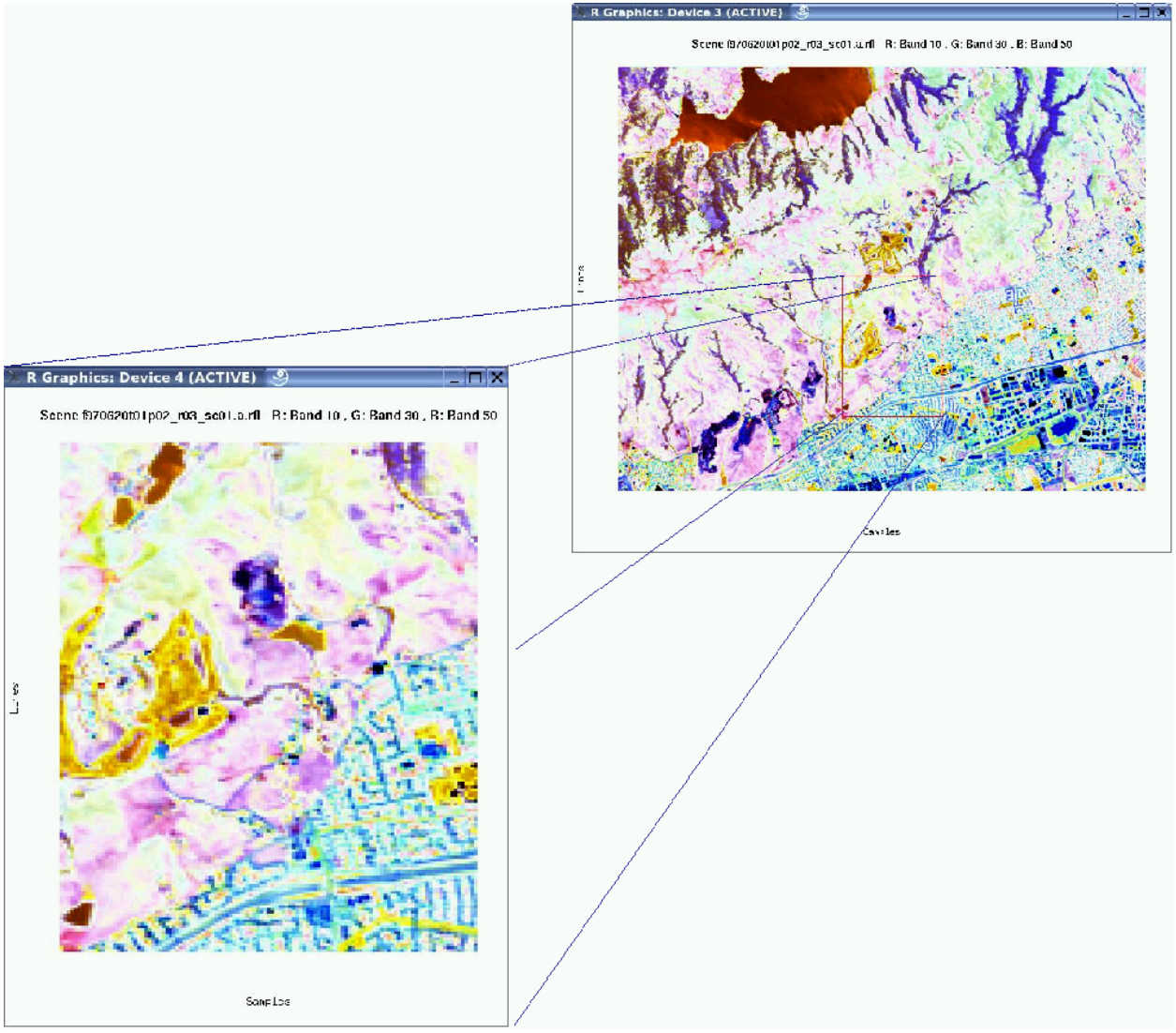
Análisis de Componentes Principales - ACP

Datos

Datos satelitales provistos por el sensor AVIRIS (Airborne Visible/Infrared Imaging Spectrometer)

- Identificación, medición y monitoreo de constituyentes de la superficie y la atmósfera terrestres basado en la absorción molecular y firma espectral de las partículas.
- 224 bands (0.4 - 2.5 μm)
- Ancho de cada banda aprox. 0.1μm

Componentes Principales



Componentes Principales

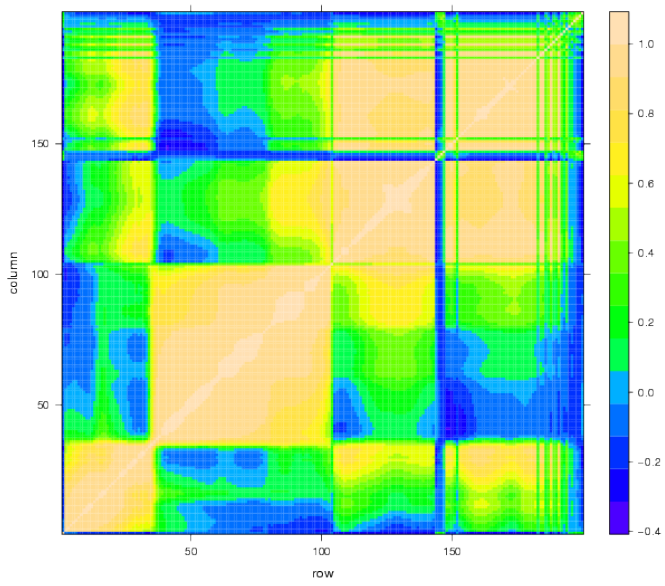
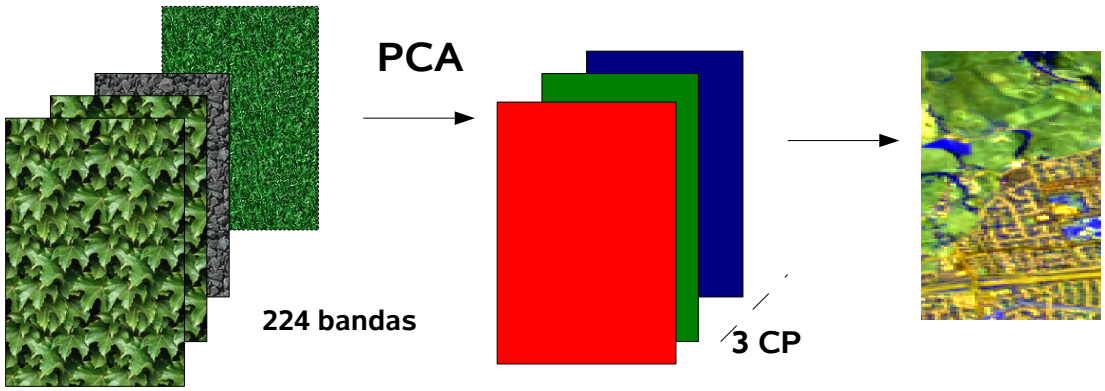


Figure: Matriz de Correlación de la imágen en estudio.

Dificultades:

- Excesiva correlación entre bandas (características) "vecinas"
- Gran volúmen de datos involucrados

Ejemplo-Datos AVIRIS



Análisis de Componentes Principales

Problema

Encontrar un espacio de dimensión más reducida que represente adecuadamente los datos y brinde la mejor representación de la variabilidad y diversidad de los mismos.

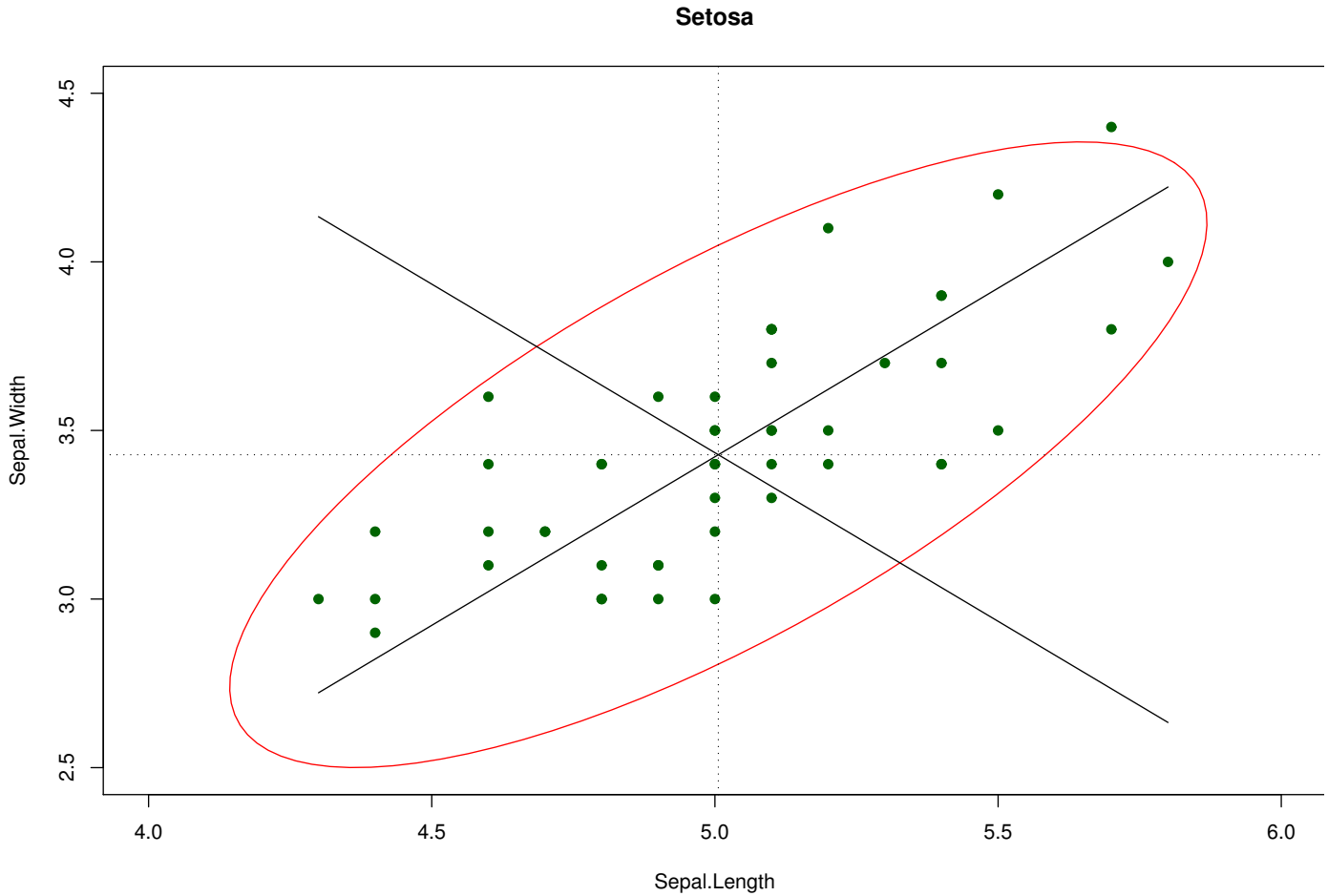
Objetivos

- Reducir dimensionalidad describiendo las p variables de una matriz X por un subconjunto (pequeño) $r < p$ de combinaciones lineales de las variables originales.
- Describir patrones de correlación entre las variables involucradas.

Análisis de Componentes Principales

- Herramienta exploratoria: técnica basada en una muestra para facilitar descripción de los datos.
- Aplicaciones:
  - Descripción e interpretación de un conjunto de datos.
  - Utilizada como técnica de pre-procesamiento en diversas aplicaciones(agrupamiento, regresión, etc)
  - Utilizada en distintas disciplinas (economía, meteorología , procesamiento de imágenes de teledetección, psicología, etc).

Enfoque Geométrico



Enfoque Geométrico

- Si las variables x_i están correlacionadas entonces, en general, la nube de puntos forma un elipsoide con centro en $\bar{\mathrm{x}}$ cuyos ejes principales no son paralelos a los ejes cartesianos.
- La dirección del eje mayor del elipsoide y la proyección de los puntos sobre esta permiten describir la orientación de la nube de puntos. Este eje minimiza las distancias ortogonales de las observaciones a una recta que pase entre ellas.
- Encontrar los ejes del elipsoide es equivalente a encontrar la matriz ortogonal A que rota los ejes de manera tal que los alinea con los ejes del elipsoide.

Pasos de un PCA

- Seleccionar las variables (descartar categóricas, etc.)
- Centrar las variables respecto a su media $\mathrm{x}_k - \bar{\mathrm{x}_k}$. Esto no cambia la estructura de la nube de puntos.
- Decidir si se van a estandarizar las variables o no. Si las variables tienen distintas unidades o magnitudes muy disímiles deben estandarizarse.
- Determinar el número de componentes que se desean retener.
- Si es necesario rotar componentes para mejorar intrepretabilidad
- Interpretar resultados.

- Primer componente principal es la dimensión en la cual las variables están más dispersas(varianza máxima).
- Segunda componente principal combinación lineal con máxima varianza con dirección ortogonal a la primer componente.
- ...
- Estas nuevas variables (PC) son no correlacionadas.

En lo que resta: Sea X, $n \times p$ matriz de observaciones. Supondremos variables $\mathrm{x}_1, \ldots, \mathrm{x}_p$ centradas respecto a sus medias.

PCA- Enfoque algebraico

- Encontrar un subespacio de dimensión $r < p$ tal que la proyección de los puntos sobre el mismo preserve la estructura (posiciones relativas) con la menor distorsión posible.
- Se busca una combinación lineal
  $\mathrm{z}_1 = a_{11}\mathrm{x}_1 + a_{12}\mathrm{x}_2 + ... + a_{1p}\mathrm{x}_p = a_1'\mathrm{x}$ de las variables originales que tenga varianza máxima.

PCA- Enfoque algebraico

Los valores de la primer componente en los n individuos se representa por el vector

$$z_1 = Xa_1$$

- $\bar{z_1} = 0$ (variables originales centradas respecto a su media)
- $\text{var}(z_1) = \frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X a_1 = a_1' S a_1$.

Para maximizar esa varianza, pidiendo además que $a_1' a_1 = 1$, se debe resolver:
$Sa_1 = \lambda_1 a_1$
Luego a_1 y λ_1 son un autovector de S y su autovalor correspondiente. Además $\lambda = \text{var}(z_1)$ y a_1 define los coeficientes de cada variable en la primer componente principal.

Pasos PCA

- Resto de las componentes se obtiene calculando los autovectores y autovalores de S (o R).
- Se ordenan los autovalores de mayor a menor, $\lambda_1 \geq \lambda_2 \geq ...$, la k-ésima PC es $z_k = a_k' x$, a_k autovector correspondiente a λ_k
- Los a_i son ortogonales
- En algunos casos es conveniente usar la matriz de correlación R en lugar de S: si las varianzas difieren substancialmente o las unidades de medición son inconmensurables las componentes de S serán dominadas por las variables con mayor varianza.

Propiedades

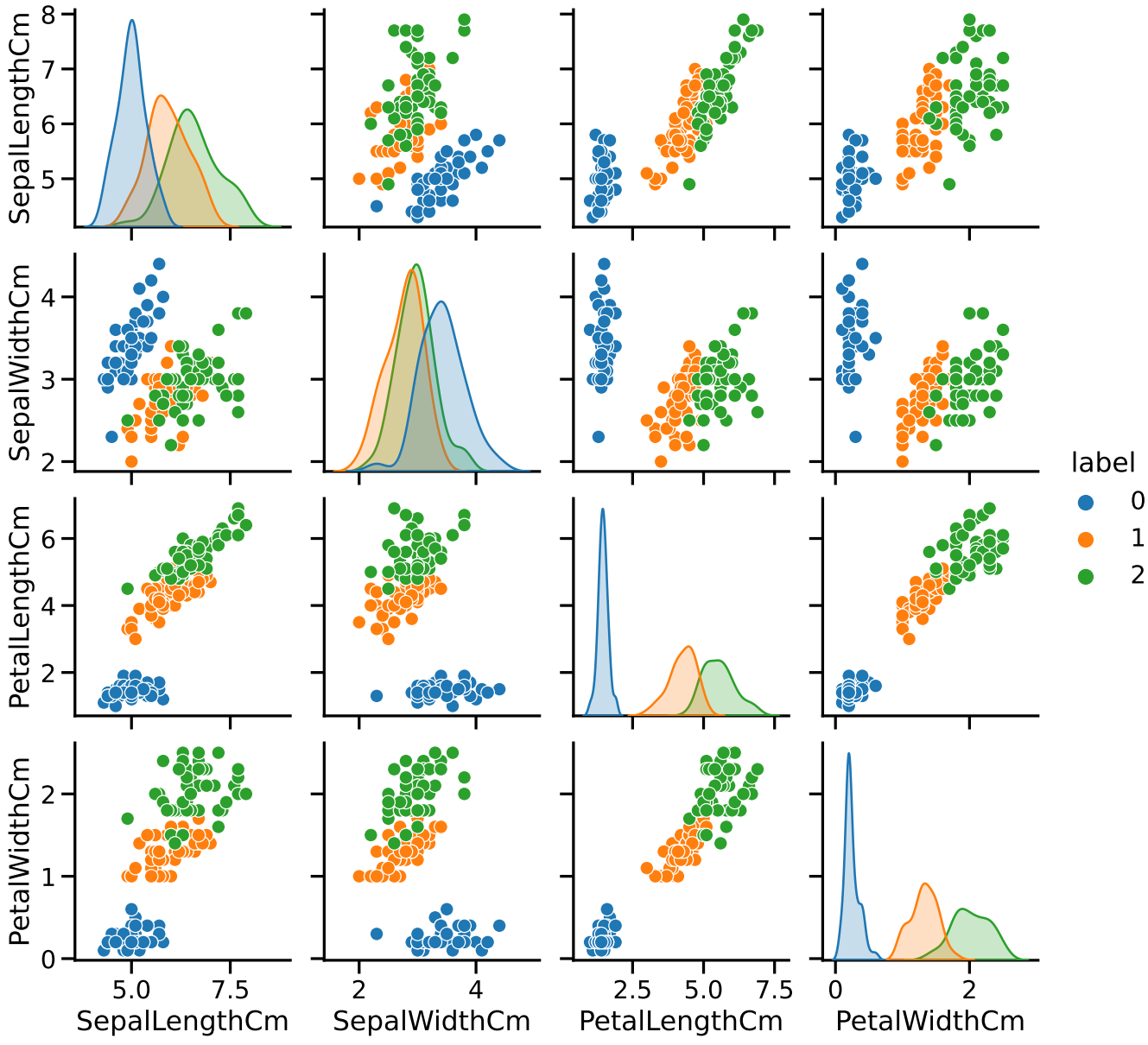
- $\sum_{i=1}^{p} \mathrm{var}(\mathrm{z}_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \mathrm{var}(\mathrm{x}_i)$
- Proporción de variabilidad explicada por la componente z_k es $\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$
- $\mathrm{cov}(\mathrm{z_i}, \mathrm{x_j}) = \lambda_\mathrm{i} \mathrm{a_{ij}}$, $\mathrm{cor}(\mathrm{z_i}, \mathrm{x_j}) = \frac{\lambda_\mathrm{i} \mathrm{a_{ij}}}{\sqrt{\lambda_\mathrm{i} \mathrm{s_j^2}}}$

- Datos: Usaremos el conjunto `iris`
- $p = 5$ `Variables: largo y ancho de sépalo (Sepal.Length, Sepal. Width), largo y ancho de pétalo (Petal.Length, Petal.Width) para flores de tres especies de iris (Species): setosa, versicolor y virginica.`
- n `= 150 individuos (50 por cada especie)`

Estadística Descriptiva

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

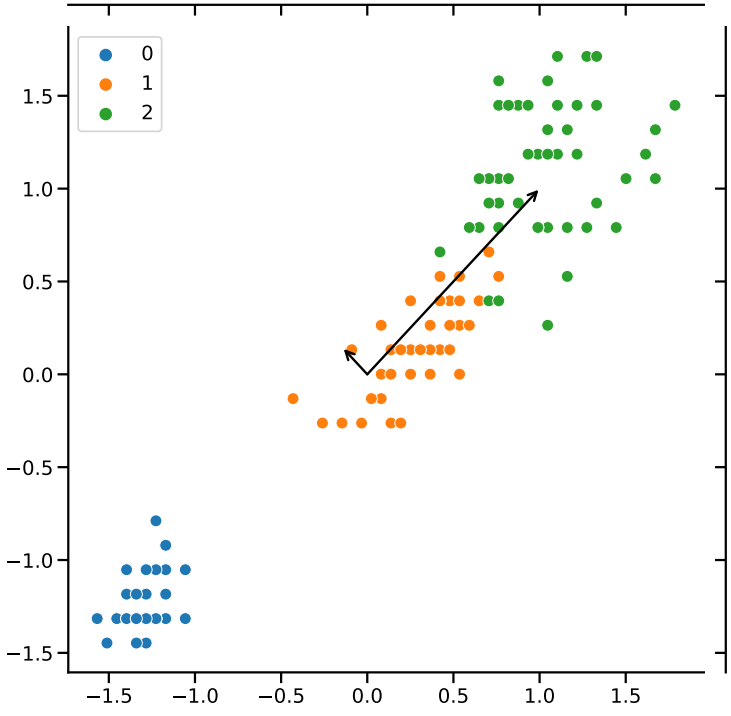
Ejemplo iris:



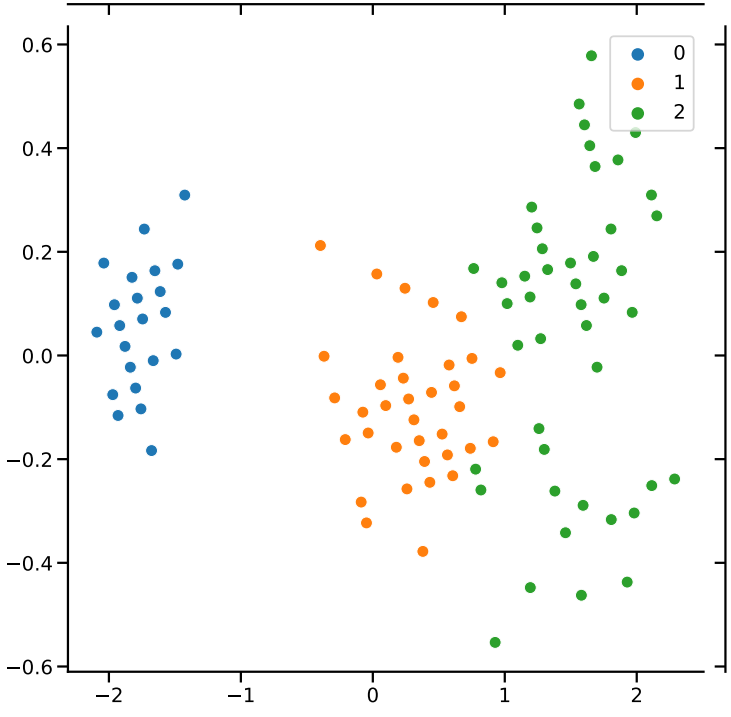
Ejemplo iris

Supongamos que el dataset consiste de solo dos variables, PetalLengths y PetalWidth

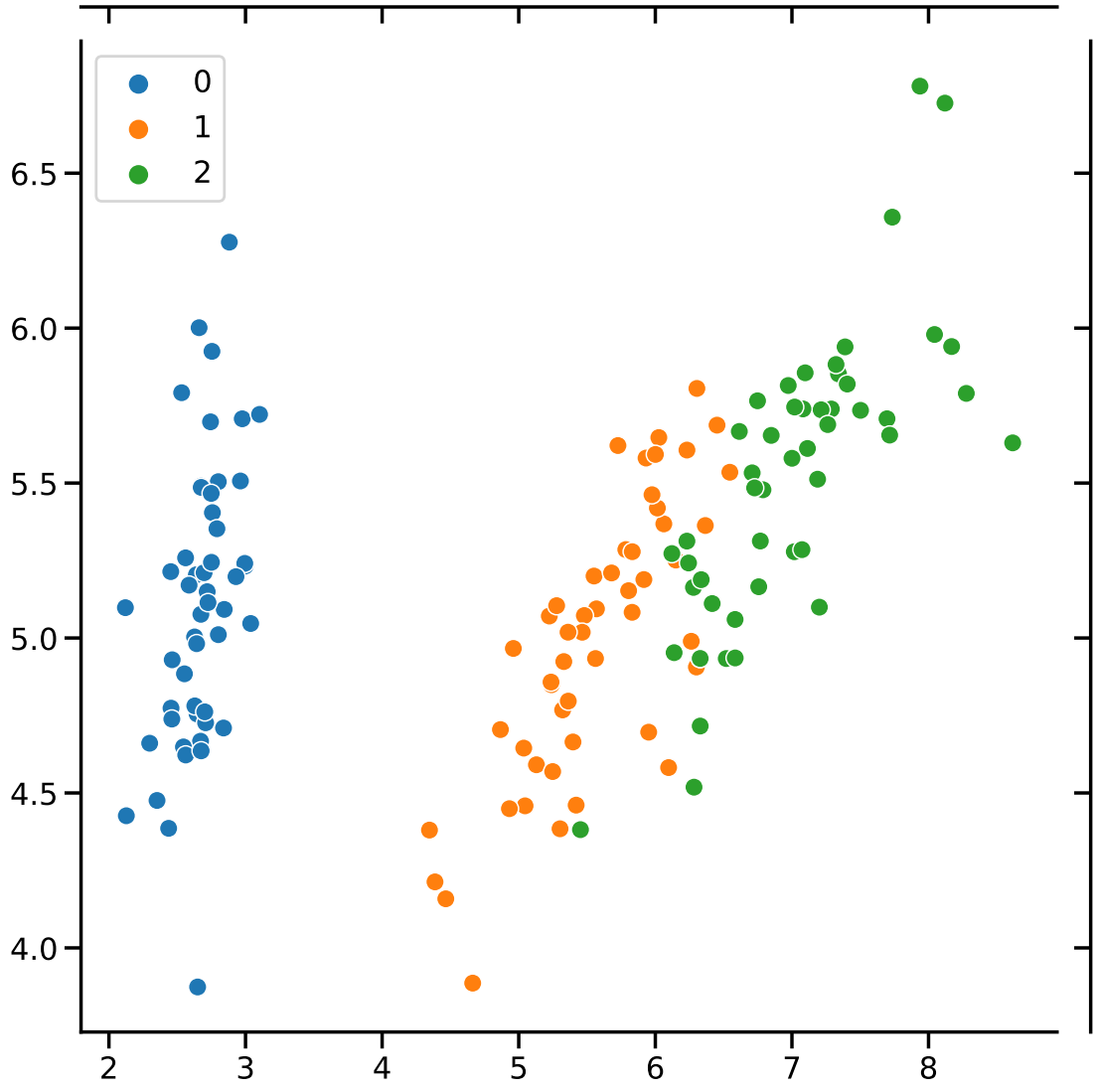
Datos originales



Proyección en Componentes principales



Ejemplo iris: ACP dataset completo



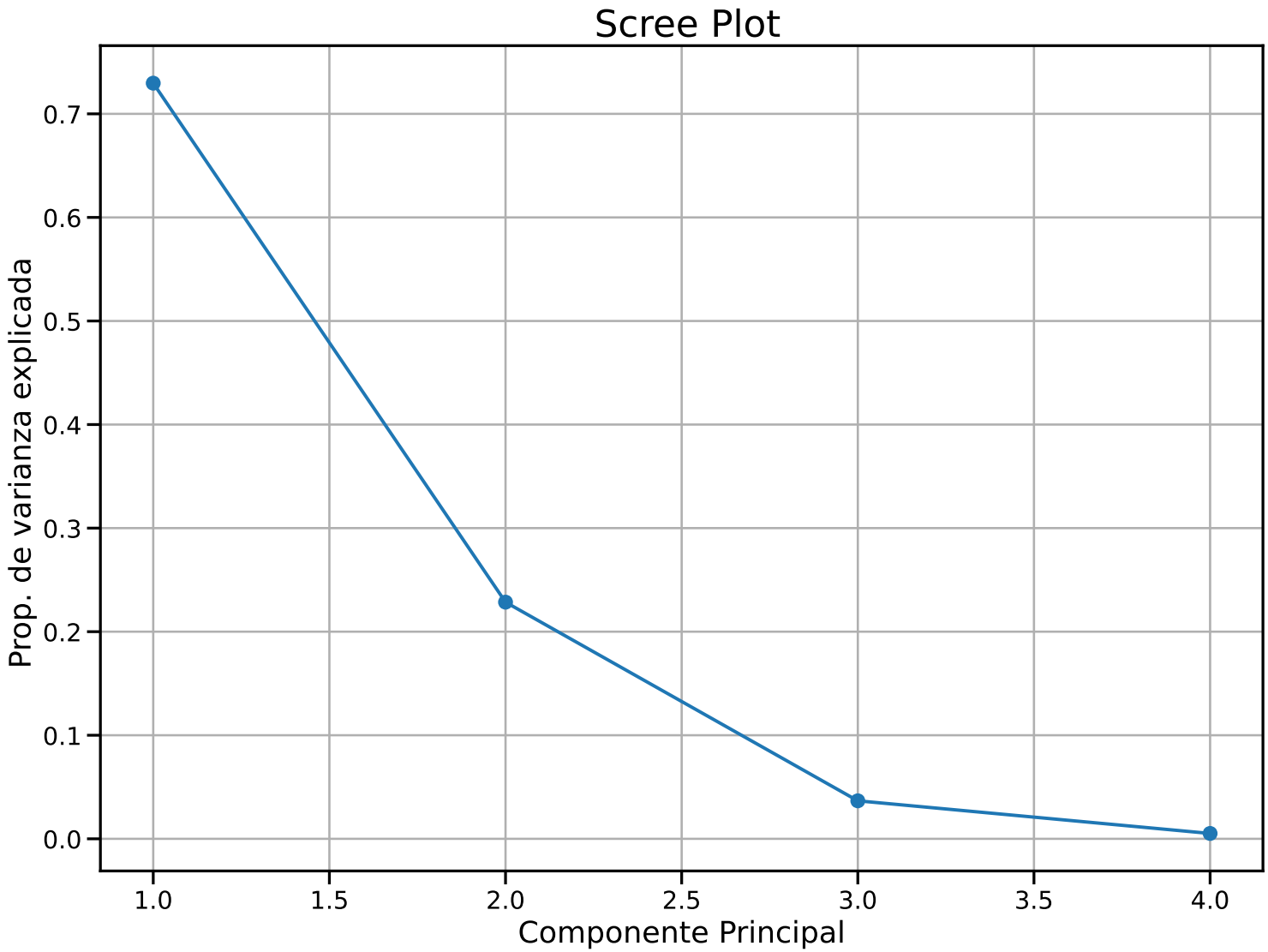
ACP: ¿Cuántas componentes seleccionar?

- Graficar λ_i vs i y buscar el corte (codo) entre autovalores "grandes" y "pequeños" (scree plot)
- Seleccionar las componentes necesarias hasta lograr una proporción determinada de la varianza (80%, 90%).
- Seleccionar las componentes cuyos autovalores sean mayores que el promedio de los mismos $\sum_{i=1}^{p} \lambda_i / p$.

En el ejemplo,

```
Componentes Principales
[[ 0.52106591 -0.26934744  0.5804131   0.56485654]
 [ 0.37741762  0.92329566  0.02449161  0.06694199]
 [-0.71956635  0.24438178  0.14212637  0.63427274]
 [-0.26128628  0.12350962  0.80144925 -0.52359713]]
Prop. de varianza explicada
[0.72962445 0.22850762 0.03668922 0.00517871]
```

Screeplot


Scree Plot
Prop. de varianza explicada
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Componente Principal

Ejemplo iris: cargas de variables en nuevas componentes

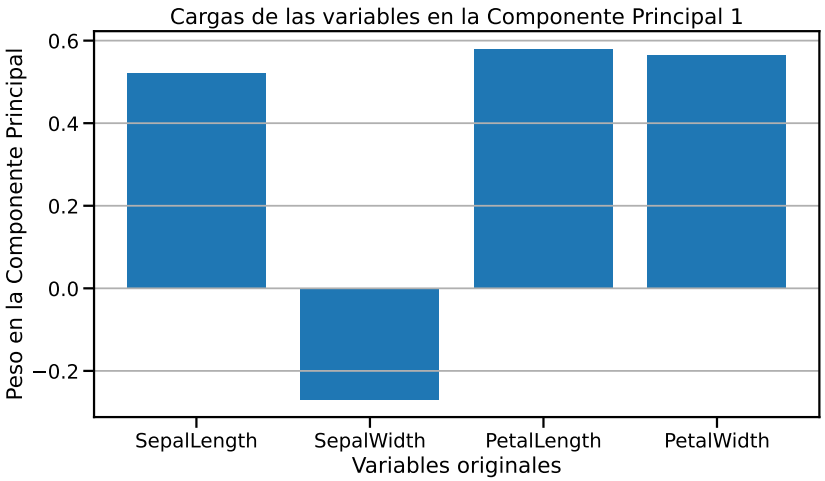


Figure: Cargas de cuatro variables originales en primer componente

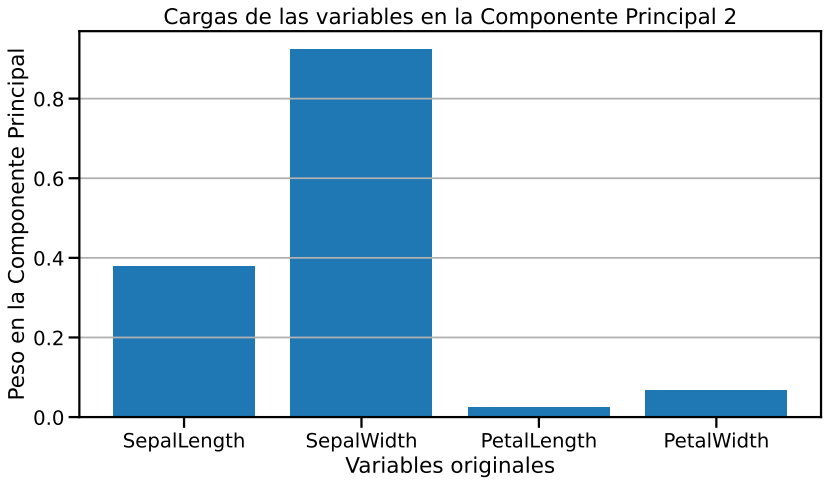


Figure: Cargas de cuatro variables originales en segunda componente

El Análisis de Componentes Principales (ACP) permite:

- Analizar los individuos: Hacer particiones entre individuos al detectar similaridades (distancia euclídea) entre ellos respecto a algunas variables o combinaciones de las mismas
- Analizar las variables: Se encuentran relaciones lineales entre las variables por medio de la descomposición de la matriz de correlacion R (o bien S).
- Pueden describirse grupos de individuos por las variables

Contenidos

Cluster o conglomerados

Agrupamiento

- Consiste en ordenar objetos en grupos de forma que el grado de asociación/similitud entre miembros del mismo cluster (clase) sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clusters.
- Permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori pero que pueden ser útiles una vez que se han encontrado.

Agrupamiento - Objetivos

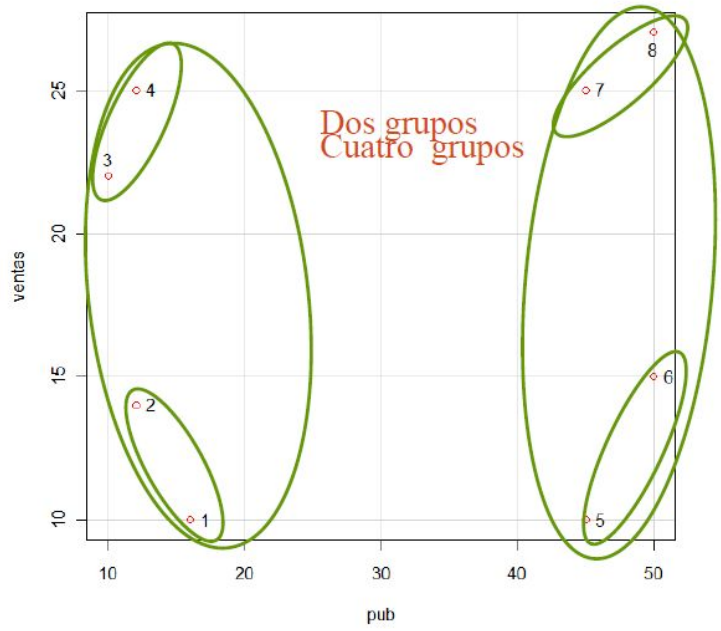
- Buscar patrones en un conjunto de datos agrupando los individuos (variables) en conglomerados (clusters).
- Agrupamiento óptimo:
  - elementos en cada grupo similares entre sí
  - grupos no similares entre sí
- "Similaridad"
  - Alguna medida de distancia,
  - Comparación entre la variabilidad dentro del cluster y entre los clusters
- Econtrar agrupamiento "natural" de los datos.

Agrupamiento

- Otros nombres: Reconocimiento de patrones (métodos de clasificación no supervisada), taxonomía numérica.
- Aplicaciones: Medicina, psiquiatría, geología, ingeniería, teledetección, meteorología, estudios de mercado, etc.

Ejemplo

¿Cuántos grupos?



Agrupamiento

Tipos de algoritmos

- Por partición de datos (ej: **k-means**)
- Métodos basados en densidad (ej:**DBSCAN**)
- Métodos Jerárquicos: (ej: Ward)
- Métodos basados en distribuciones (ej: **mezcla de gaussianas**, MeanShift)

Agrupamiento

Partición de datos

Se dividen los individuos en g grupos (g prefijado) internamente homogéneos:

- Cada individuo pertenece solamente a un grupo
- Todos los individuos quedan clasificados

Métodos basados en densidad

- identifica distintos grupos en los datos basándose en la idea que un cluster es una región de alta densidad de puntos
- regiones de alta densidad de puntos están separadas por regiones de baja densidad de puntos

Agrupamiento

Métodos Jerárquicos

- Se estructuran los individuos(variables) en forma jerárquica por su similitud.
- Los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores.
- Se comienza con n clusters (1 por individuo) y se termina con un cluster que contiene los n individuos (o viceversa).
- En cada paso, se absorben un individuo o un cluster de individuos en otro cluster.
- Definen la estructura de asociación en cadena que puede existir entre los elementos.

Agrupamiento

Métodos basados en distribuciones

- Asume que los datos siguen algún tipo de distribución probabilística.
- un dato forma parte de un grupo según la probabilidad que pertenezca a ese grupo
- hay un punto central, a medida que aumenta la distancia entre ese pnto central y un dato disminuye la probabilidad que ese dato forme parte de ese grupo

Método de Partición: K-means

Número k de clusters que se desea obtener se define previamente.

1. Seleccionar k observaciones como centro de los grupos iniciales (semillas):
2. Calcular distancias euclídeas de cada observación a los centroides de los k clusters. Asignar cada elemento al cluster más próximo. Recalcular centroides.
3. Una vez asignados todas las observaciones a los clusters se decide en base a un criterio de optimalidad o convergencia si deben realocarse las observaciones y actualizarse los centroides.
4. Repetir 2 y 3 hasta que no se pueda hacer ninguna mejora, o hasta satisfacer un número máximo de iteraciones.

K-means

Selección de semillas (baricentros) iniciales:

- Seleccionar k observaciones aleatoriamente que estén mutuamente separadas por una distancia $> r$
- Seleccionar las primeras k observaciones del conjunto de datos que estén mutuamente apartadas por una distancia $> r$
- Seleccionar k observaciones del conjunto de datos que estén mutuamente más apartadas
- Seleccionar los k centroides de la solución de k clusters de un método jerárquico.

Resultados sensibles a la elección de las semillas iniciales y a la presencia de outliers.

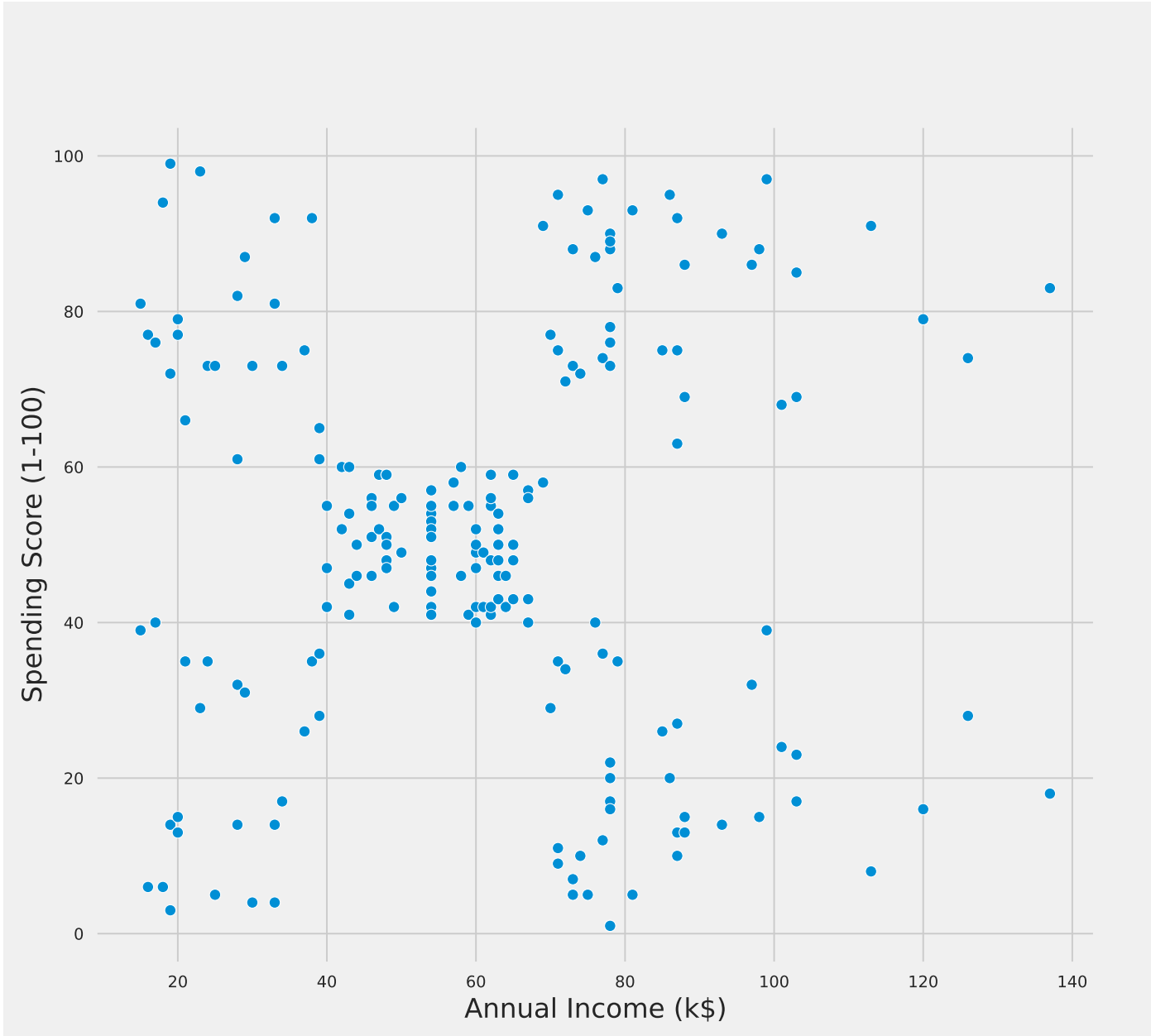
Estadística Descriptiva

Ejemplo: Datos consumo

Ilustraremos con la base de datos "MallCustomers.csv" descargada de `https://www.kaggle.com/datasets/`donde se registraron datos de ingresos y consumos anuales de un grupo de personas basados en género y edad.

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Datos consumo



Algoritmo K-means

Criterio de convergencia - Suma de cuadrados dentro

Minimizar las distancias (euclideas) al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo, esto es, minimizar:

$$W = \sum_{g=1}^{k} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)(x_{ig} - \bar{x}_g)^t$$

n_g número de observaciones en el grupo g, $\bar{x}_g$= media de ese grupo

- Minimizando distancias de todas las variables en los grupos $\Rightarrow$ grupos más homogéneos.
- No es invariante ante cambios de escalas $\Rightarrow$ conviene estandarizar variables si están en distintas unidades

Algoritmo K-means

Este algoritmo busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro.

1. Fijado el número de grupos y seleccionados los centros,
2. comprobar si moviendo algún elemento se reduce W
3. Si la respuesta es positiva $\Rightarrow$ mover elemento, recalcular medias (centroides) de los dos grupos afectados por el cambio y volver al paso anterior.
4. Si no es posible reducir W o se excedió un número prefijado de iteraciones $\Rightarrow$ terminar.

Observaciones

- Resultado afectado por semillas iniciales y asignación inicial de elementos a grupos
- Se sugiere implementarlo varias veces desde distintas semillas iniciales

Algoritmo K-means - ¿Como determinar número de clusters?

1. No existe un criterio óptimo
2. Puede seleccionarse k = número de clusters final de algún método jerárquico
3. Criterio empírico: Calcular la diferencia entre la SCD con g y $g+1$ grupos, analizando la reducción de variabilidad relativa luego de un agrupamiento adicional:

$$F = \frac{\mathrm{W(g) - W(g+1)}}{\mathrm{W(g+1)/(n-g-1)}}$$

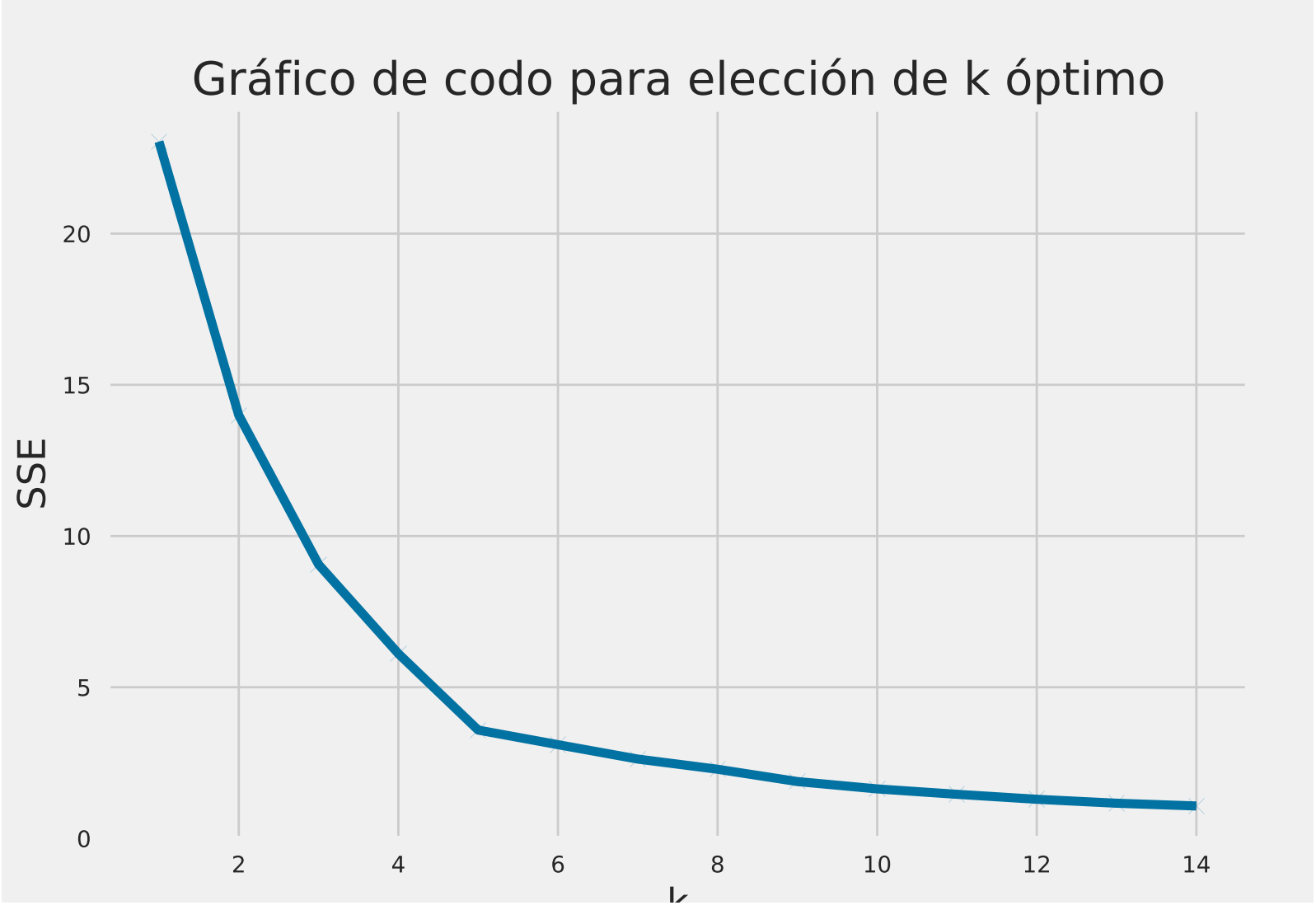
   Se compara con una distribución F con $p, p(n-g-1)$ grados de libertad. Si el cociente $F > 10$, se sugiere usar $g+1$ grupos (Hartigan (1975))
4. Gráfico de inercias en función de g (cantidad de grupos),
5. Coeficiente de silhoutte (siluetas)

Gráfico de inercias

$$\text{Inercia} = \sum_{i=1}^{g} \sum_{x \in C_i} d^2(x, \bar{x}_i)$$

- La inercia es la suma de distancias cuadradas dentro de cada cluster en la partición final
- Es una medida de cuan coherentes son los grupos
- Si se grafica la inercia en función de g, el número de grupos, se considera que el número de grupos más apropiado ocurre cuando se desacelera la reducción de la inercia.

Gráfico de inercias



Análisis de siluetas

Para cada observación $x_i \in C_k$ se calculan los indices :

- similaridad promedio $a(i) = \frac{1}{|C_k|-1} \sum_{j \in C_k, i \neq j} d(i,j)$
- disimilaridad mínima promedio
  $b(i) = \min_{k \neq l} \frac{1}{|C_l|} \sum_{j \in C_l} d(i,j)$

Se definen

$$s(i) = \begin{cases} 1 - a(i)/b(i) & si\ a(i) < b(i) \\ 0 & si\ a(i) = b(i) \\ b(i)/a(i) - 1 & si\ a(i) > b(i) \end{cases}$$

y

$$SC = \frac{1}{n_g} \sum_{i=1}^{n_g} s(i)$$

Métrica de uso interno (elección de medida de distancia, algoritmo de agrupamiento o de número de grupos) que indica cuán similar es un elemento a su propio grupo (cohesión) en comparación con otros grupos (separación)

Análisis de siluetas

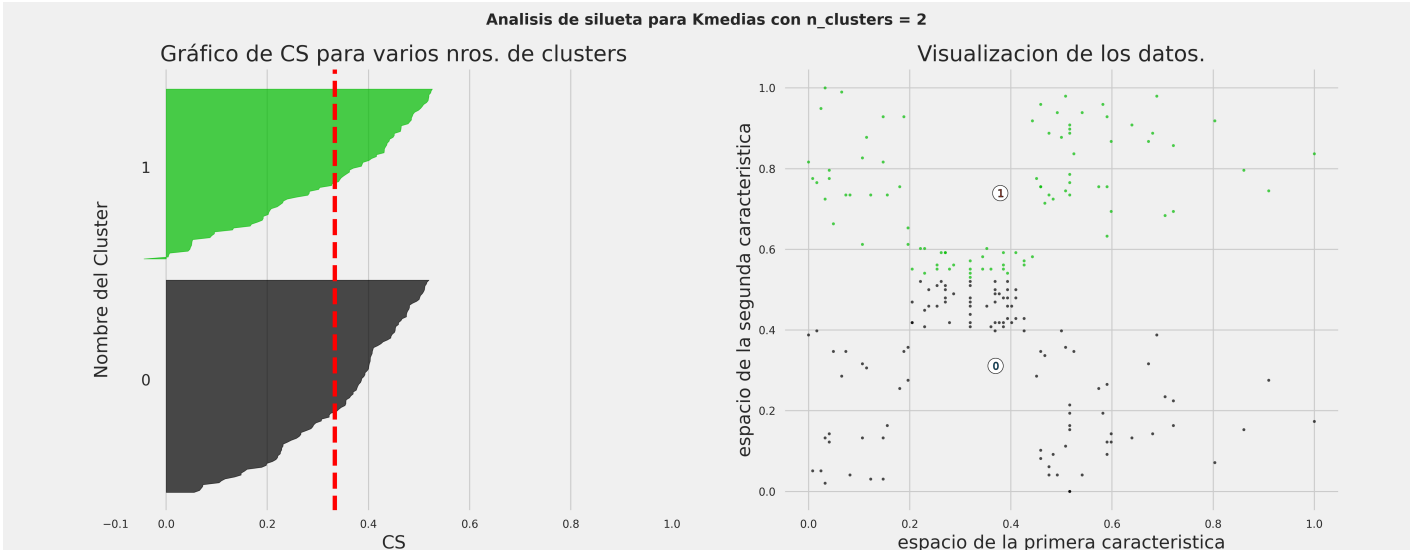
Observaciones

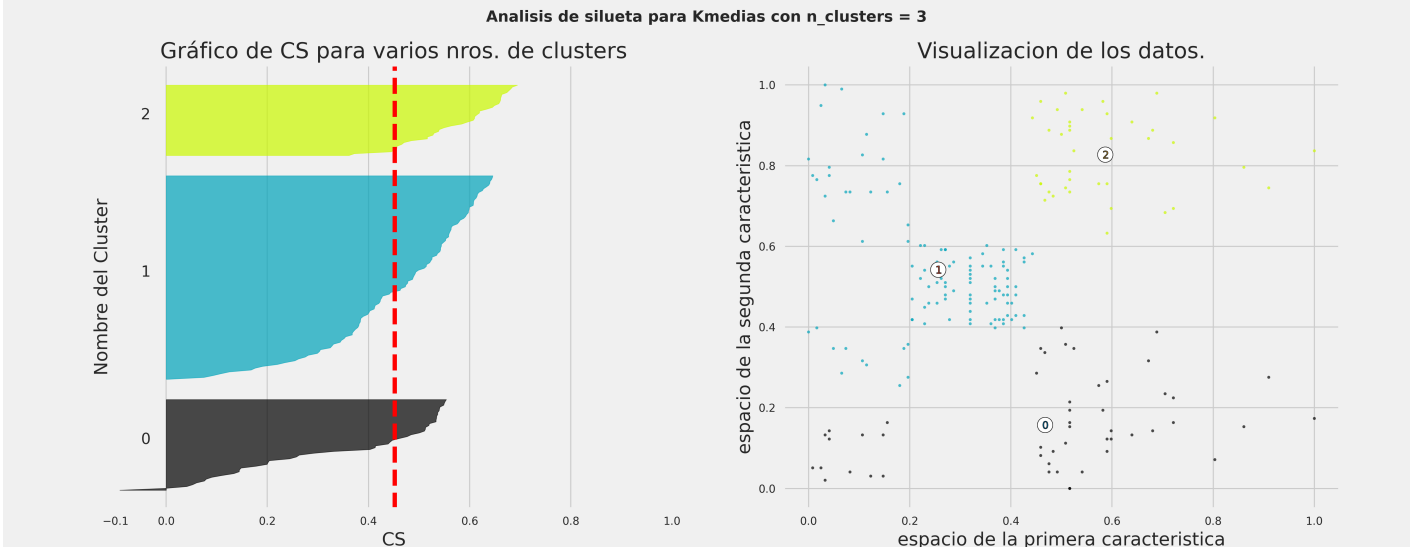
- $-1 < s(i) < 1$
- Coeficientes $s(i)$:
  - cercanos a 1 $\Rightarrow$ la muestra está lejos de los clusters vecinos
  - iguales o muy cercanos a 0 $\Rightarrow$ la muestra está muy cerca del borde de decisión entre clusters.
  - $< 0 \Rightarrow$ puntos asignados al cluster equivocado

Rousseew(1987) propuso la siguiente interpretación del coeficiente SC:

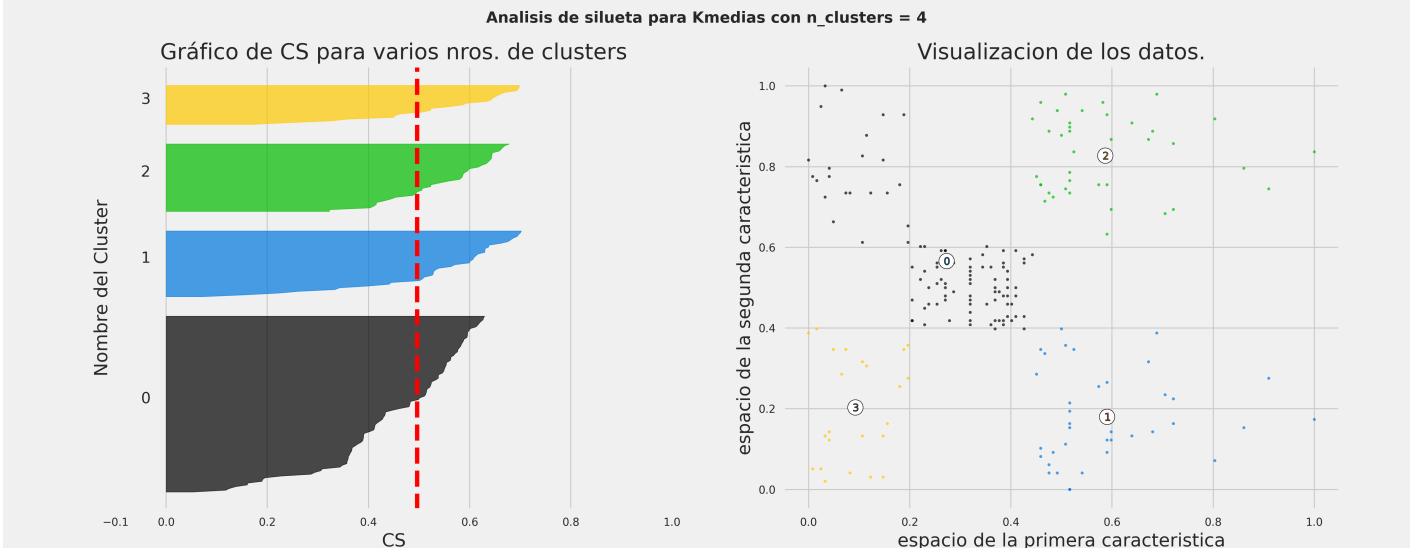
- $0.71 - 1$: estructura fuerte.
- $0.51 - 0.7$: se encontró una estructura razonable.
- $0.26 - 0.5$: estructura débil y podría ser artificial.
- menor a 0.25: sin estructura sustancial.

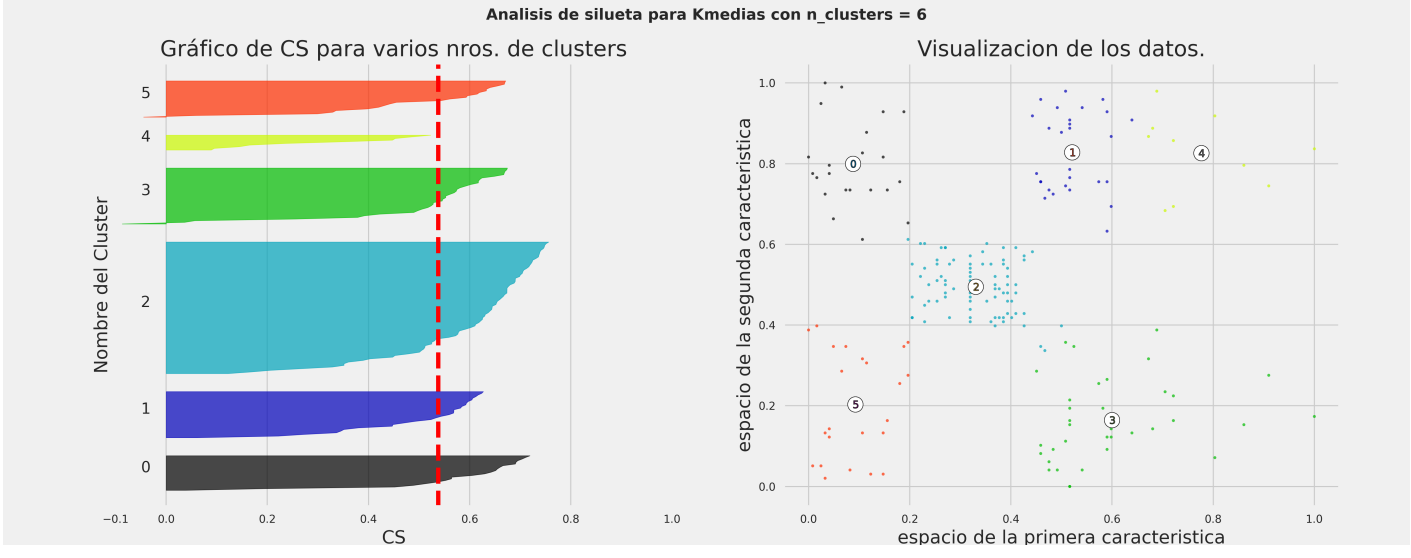
Análisis de siluetas



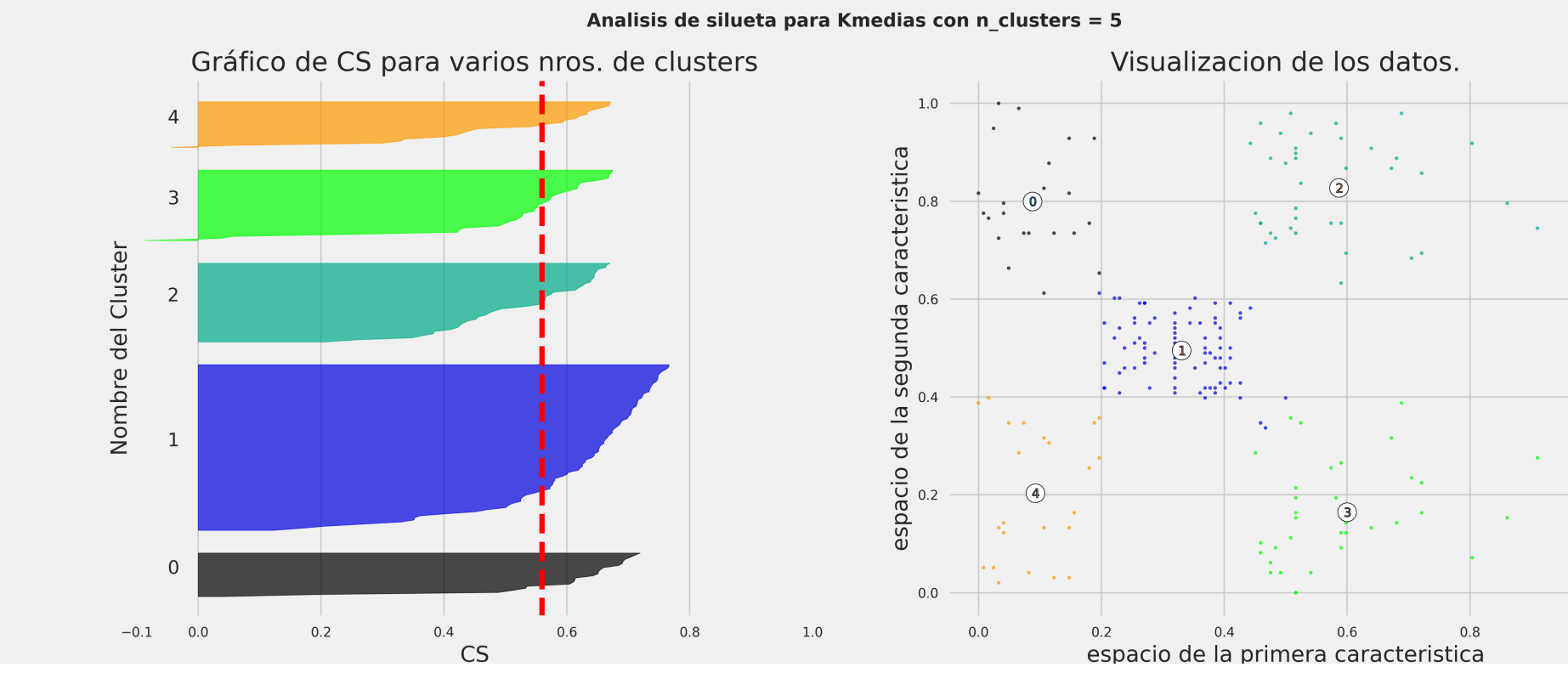


Análisis de siluetas

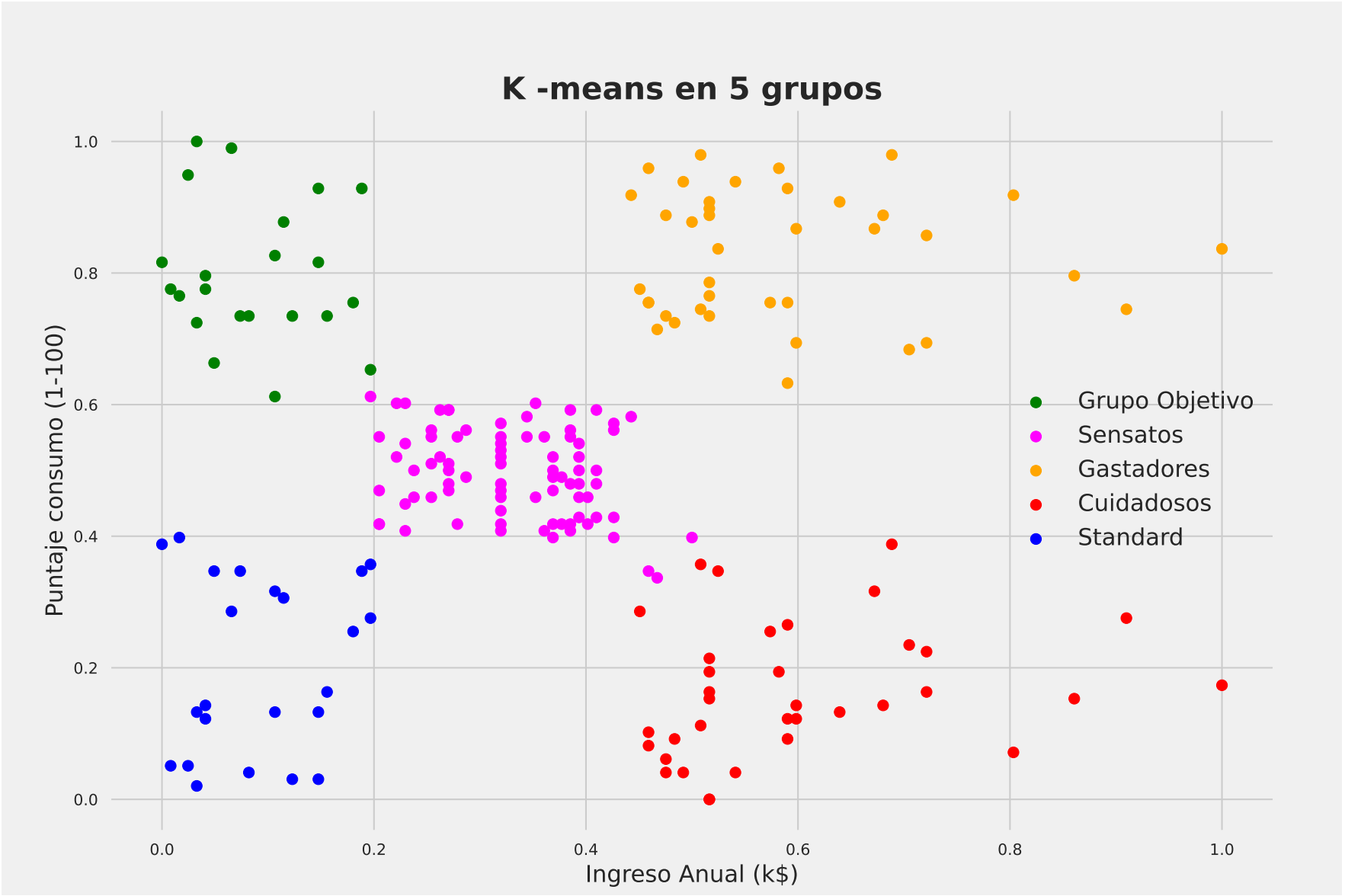




Análisis de siluetas



Ejemplo , k = 5


K -means en 5 grupos
Puntaje consumo (1-100)
Ingreso Anual (k$)
0.0
0.2
0.4
0.6
0.8
1.0
Grupo Objetivo
Sensatos
Gastadores
Cuidadosos
Standard

Métodos basados en densidades

DBSCAN

Density Based Spatial Clusterig of Applications with Noise (DBSCAN)

- Clusters: regiones densas en el espacio de datos, separadas por regiones de baja densidad de puntos.
- Idea básica: Para que un punto pertenzca a un cluster , un entorno de un radio dado, centrado en ese punto, tiene que contener una cantidad mínima de puntos
- Puede encontrar clusters de diferentes formas y tamaños , detecta ruido y puntos outliers.

DBSCAN

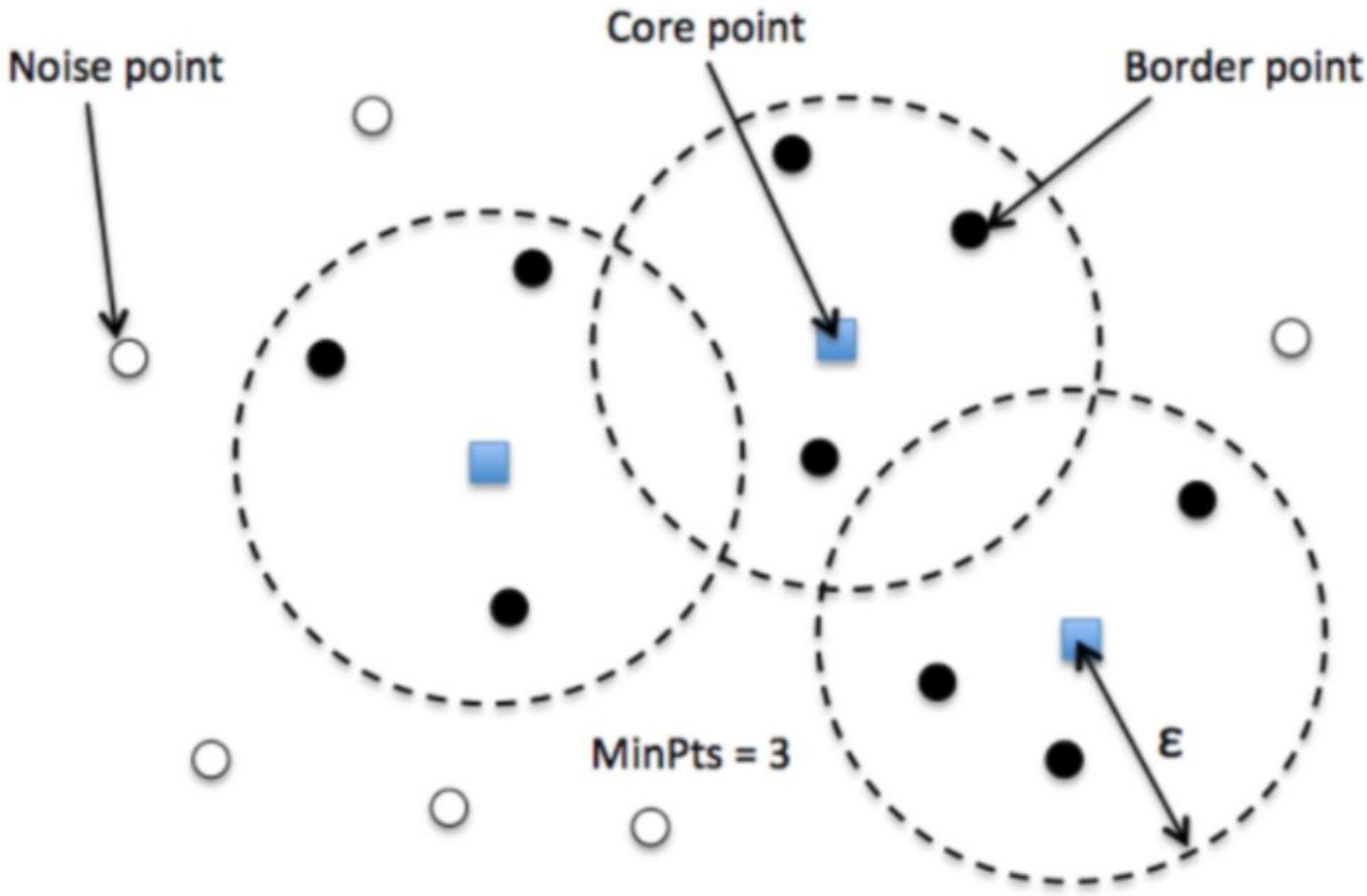
Parámetros

- ϵ (eps) = define el entorno (vecindad) de un punto: Si $d(x_i, x_j) < \epsilon \Rightarrow x_i$ y x_j son "vecinos"
- min Pts: número mínimo de vecinos (puntos) dentro de un radio ϵ (define así zonas "densas")

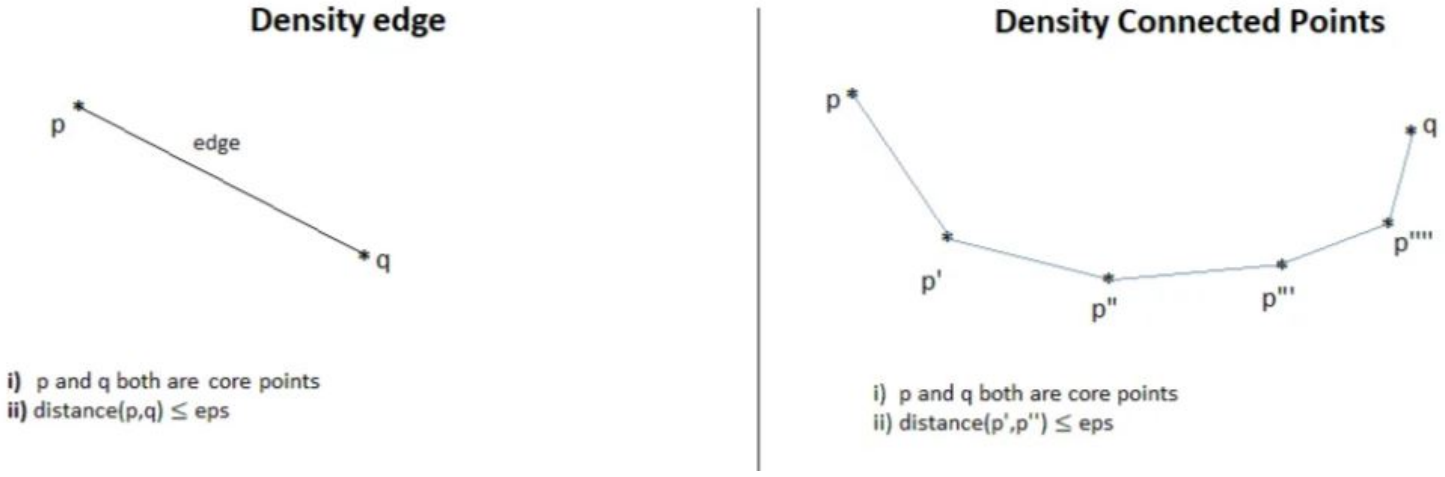
Etiquetado de los puntos:

- punto de núcleo (core): la cantidad de puntos a un radio ϵ de ese punto es mayor a min Pts.
- punto de borde: la cantidad de puntos a un radio ϵ de ese punto es menor a min Pts, pero está en el entorno de un punto de núcleo.
- ruido o outlier: punto que no es punto de núcleo ni punto de borde.

DBSCAN



DBSCAN



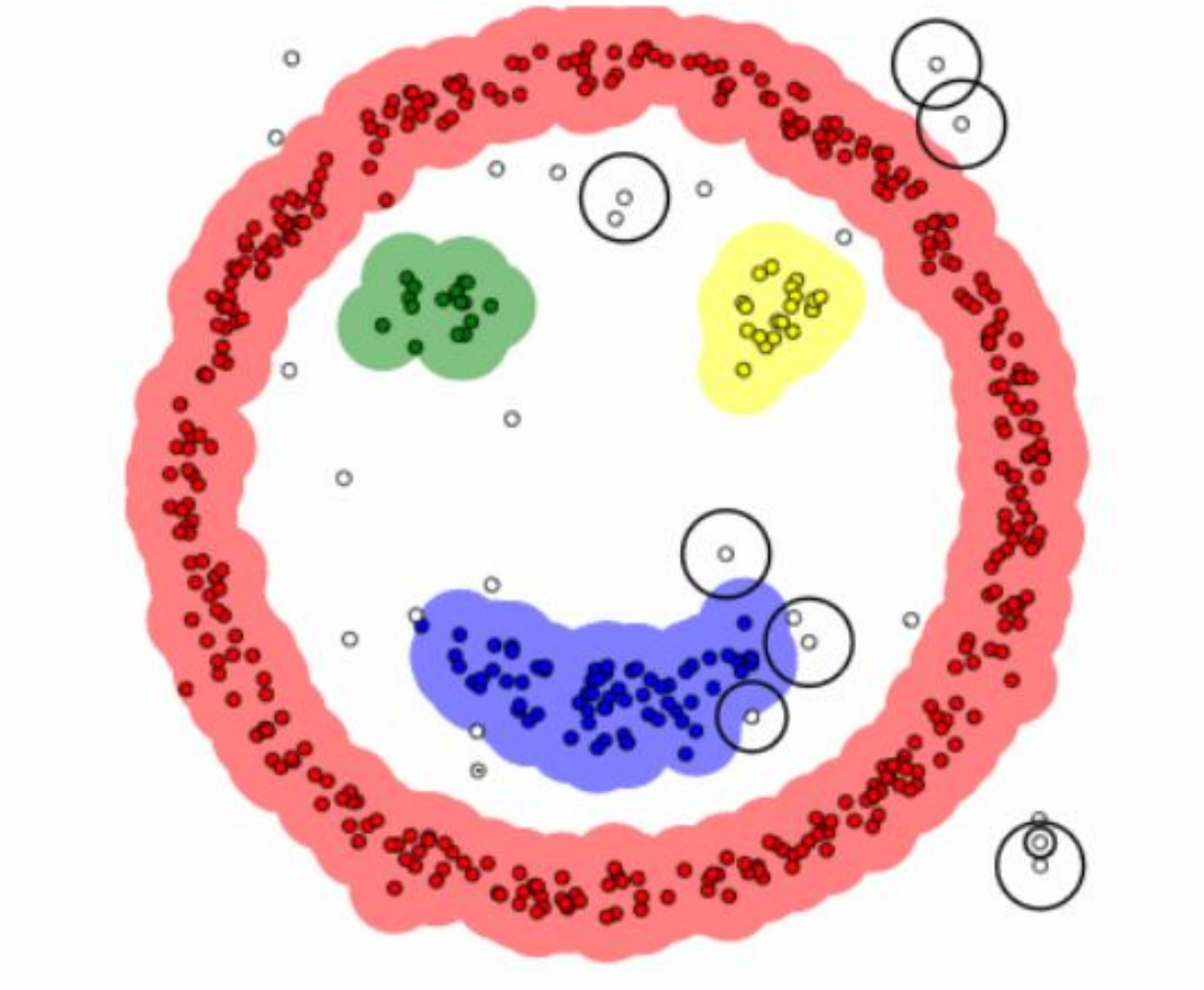
borde de densidad: si la distancia entre dos núcleos es menor a ϵ, se pueden unir esos puntos por un segmento denominado "borde de densidad"

puntos conectados por densidad: Se dice que dos puntos p y q son puntos conectados por densidades si ambos son puntos de núcleo y existe un camino formado por bordes de densidades que conectan el punto p con el punto q

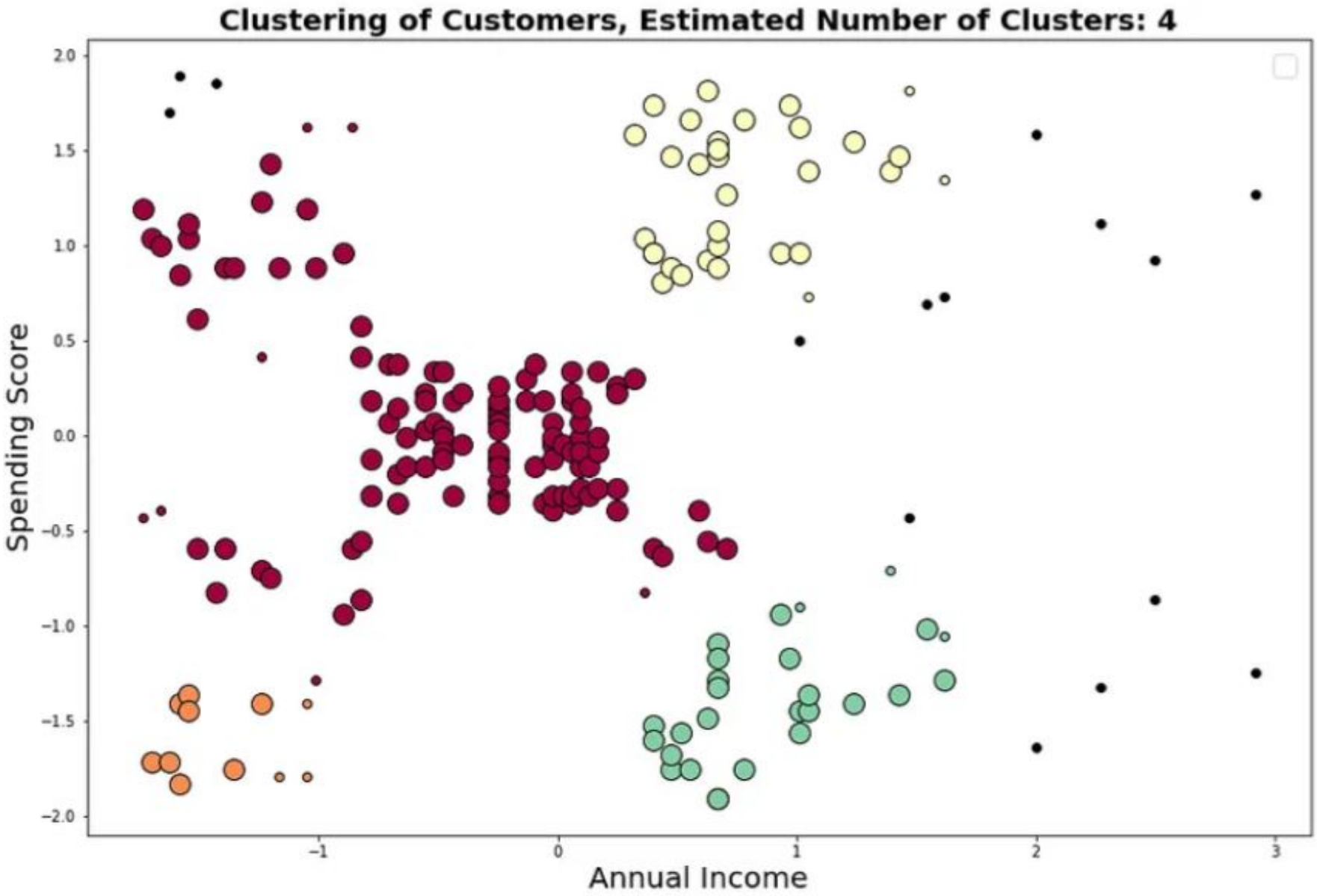
DBSCAN- Algoritmo

1. Toma un punto arbitrario y encuentra todos los puntos en el entorno de radio ϵ
2. se etiqueta el punto. Si es un punto de núcleo, comienza la formación de grupos, sino se etiqueta al punto como ruido (esta etiqueta puede modificarse más tarde, pues ese punto puede estar en el entorno de otro punto)
3. Se buscan todos los puntos conectados por densidad a ese punto núcleo y se los asigna al mismo grupo.
4. se itera sobre todos los puntos que no fueron visitados, formándose los distintos rupos. Los puntos que no pertenecen a ningún grupo son ruido o outliers.

DBSCAN



DBSCAN


Clustering of Customers, Estimated Number of Clusters: 4
Spending Score
Annual Income
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3

DBSCAN

```
Cluster 0, Avg Annual Income: 48, Avg Spending Score: 52, Count: 114
Cluster 1, Avg Annual Income: 24, Avg Spending Score: 9, Count: 11
Cluster 2, Avg Annual Income: 81, Avg Spending Score: 84, Count: 32
Cluster 3, Avg Annual Income: 84, Avg Spending Score: 14, Count: 27
```

Figure: DBSCAN, $\epsilon = 0.4$, min Pts = 5

```
Cluster 0, Avg Annual Income: 21, Avg Spending Score: 75, Count: 10
Cluster 1, Avg Annual Income: 25, Avg Spending Score: 32, Count: 5
Cluster 2, Avg Annual Income: 55, Avg Spending Score: 49, Count: 87
Cluster 3, Avg Annual Income: 79, Avg Spending Score: 84, Count: 27
Cluster 4, Avg Annual Income: 76, Avg Spending Score: 10, Count: 14
Cluster 5, Avg Annual Income: 90, Avg Spending Score: 14, Count: 7
```

Figure: DBSCAN, $\epsilon = 0.25$, min Pts = 5

```
Cluster 0, Avg Annual Income: 23, Avg Spending Score: 75, Count: 11
Cluster 1, Avg Annual Income: 55, Avg Spending Score: 49, Count: 87
Cluster 2, Avg Annual Income: 79, Avg Spending Score: 84, Count: 29
Cluster 3, Avg Annual Income: 80, Avg Spending Score: 13, Count: 22
```

Figure: DBSCAN, $\epsilon = 0.4$, min Pts = 10

```
Cluster 0, Avg Annual Income: 61, Avg Spending Score: 50, Count: 199
```

Figure: DBSCAN, $\epsilon = 0.75$, min Pts = 5

DBSCAN: elección de ϵ y Min Pts

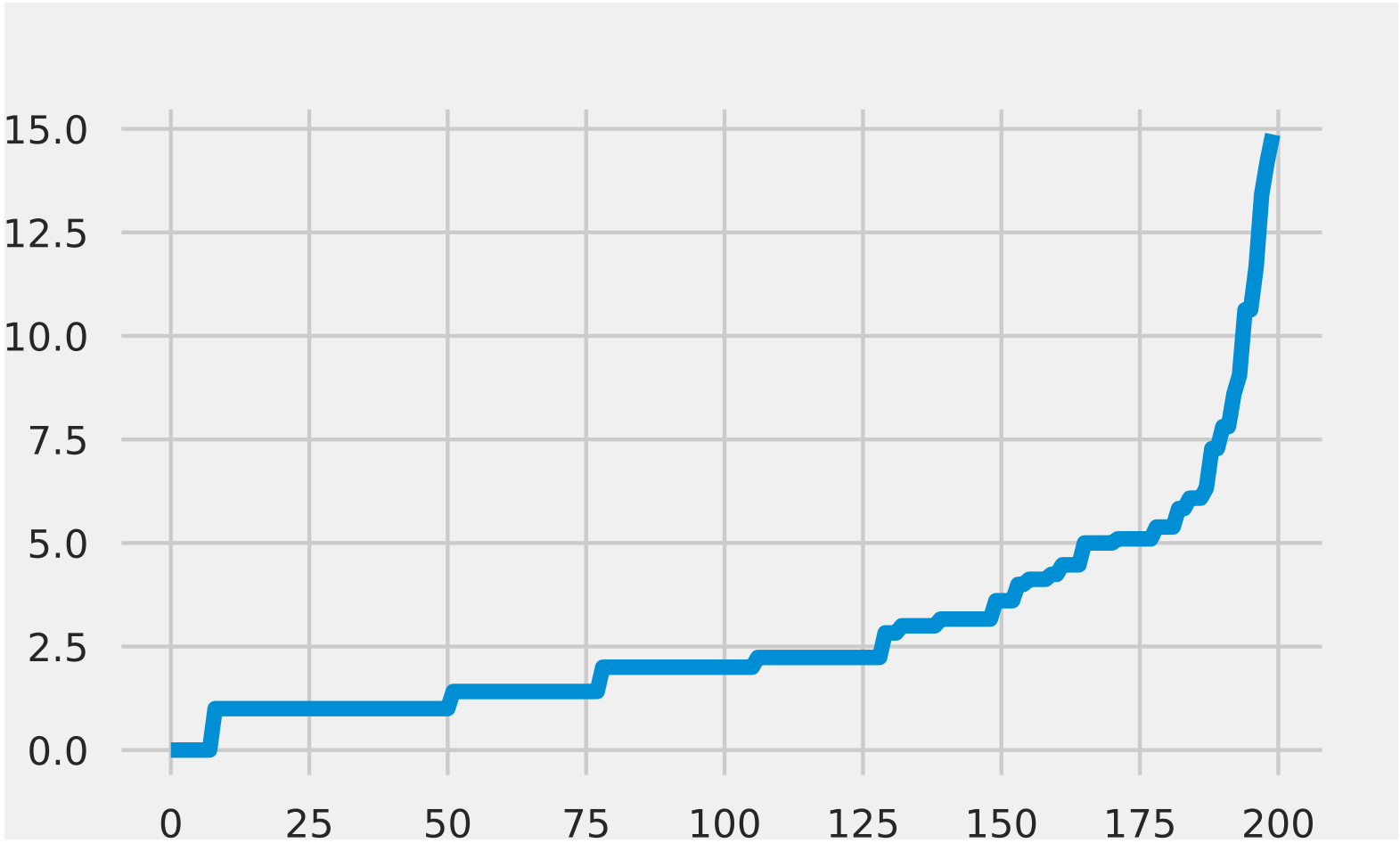
Min Pts

- Min Pts $\leq p + 1$, con p= cantidad de variables.
- si el data set es muy ruidoso, Min Pts debe ser al menos 2p

ϵ

- ϵ pequeño $\rightarrow$ muchos datos sin agrupar; ϵ muy grande $\rightarrow$ la mayoría de los datos en un mismo grupo.
- puede elegirse usando un gráfico de distancias (gráfico de codo, knee graph), representando la distancia al $k = minPts - 1$ vecino más cercano, ordenadas de mayor a menor. Se elige ϵ donde se observa un codo.
- para cada punto se detecta su vecino más cercano, se fija ϵ de manera tal que una proporción suficientemente gande de observaciones ($> 90\%$) tenga una distancia a su vecino más cercano inferior a ϵ

kNN



Kmeans vs DBSCAN